

Amendments to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

- 1 1. (currently amended): A system for grouping clusters of
2 semantically scored documents electronically stored in a data corpus, comprising:
3 a scoring module determining a score, which is assigned to at least one
4 concept that has been extracted from a plurality of electronically-stored
5 documents, wherein the score is based on at least one of a frequency of
6 occurrence of the at least one concept within at least one such document, a
7 concept weight, a structural weight, and a corpus weight;
8 a clustering module forming clusters of the documents by evaluating the
9 score for the at least one concept of each document for a best fit to the clusters
10 and assigning each document to the cluster with the best fit; and
11 a threshold module determining similarities between the documents
12 grouped into each cluster based on the center of the cluster and the scores
13 assigned to each of the at least one concepts in each such document, dynamically
14 determining a threshold for each cluster ~~based on as a function of the similarities~~
15 ~~between the documents grouped into the cluster and a center of the cluster, and~~
16 identifying and reassigning those documents having the similarities falling outside
17 the threshold.
- 1 2. (original): A system according to Claim 1, further comprising:
2 the scoring module calculating the score as a function of a summation of
3 at least one of the frequency of occurrence, the concept weight, the structural
4 weight, and the corpus weight of the at least one concept.
- 1 3. (original): A system according to Claim 2, further comprising:

2 a compression module compressing the score through logarithmic
3 compression.

1 4. (original): A system according to Claim 1, further comprising:
2 the scoring module calculating the concept weight as a function of a
3 number of terms comprising the at least one concept.

1 5. (original): A system according to Claim 1, further comprising:
2 the scoring module calculating the structural weight as a function of a
3 location of the at least one concept within the at least one such document.

1 6. (original): A system according to Claim 1, further comprising:
2 the scoring module calculating the corpus weight as a function of a
3 reference count of the at least one concept over the plurality of documents.

1 7. (currently amended): A system according to Claim 1, further
2 comprising:
3 the scoring module forming the score assigned to the at least one concept
4 to a normalized score vector for each such document, determining [[a]] each such
5 similarity between the normalized score vector for each such document as an
6 inner product of each normalized score vector, and applying the similarity to the
7 best fit criterion.

1 8. (original): A system according to Claim 1, further comprising:
2 the clustering module evaluating a set of candidate seed documents
3 selected from the plurality of documents, identifying a set of seed documents by
4 applying the score for the at least one concept to a best fit criterion for each such
5 candidate seed document, and basing the best fit criterion on the score of each
6 such seed document.

1 9. (currently amended): A method for grouping clusters of
2 semantically scored documents electronically stored in a data corpus, comprising:

3 determining a score, which is assigned to at least one concept that has
4 been extracted from a plurality of electronically-stored documents, wherein the
5 score is based on at least one of a frequency of occurrence of the at least one
6 concept within at least one such document, a concept weight, a structural weight,
7 and a corpus weight;

8 forming logically-grouped clusters of the documents by evaluating the
9 score for the at least one concept of each document for a best fit to the clusters
10 and assigning each document to the cluster with the best fit;

11 determining similarities between the documents grouped into each cluster
12 based on the center of the cluster and the scores assigned to each of the at least
13 one concepts in each such document;

14 dynamically determining a threshold for each cluster ~~based on as a~~
15 function of the similarities between the documents grouped into the cluster and a
16 ~~center of the cluster;~~ and

17 identifying and reassigning those documents having the similarities falling
18 outside the threshold.

1 10. (original): A method according to Claim 9, further comprising:
2 calculating the score as a function of a summation of at least one of the
3 frequency of occurrence, the concept weight, the structural weight, and the corpus
4 weight of the at least one concept.

1 11. (original): A method according to Claim 10, further comprising:
2 compressing the score through logarithmic compression.

1 12. (original): A method according to Claim 9, further comprising:
2 calculating the concept weight as a function of a number of terms
3 comprising the at least one concept.

1 13. (original): A method according to Claim 9, further comprising:
2 calculating the structural weight as a function of a location of the at least
3 one concept within the at least one such document.

1 14. (original): A method according to Claim 9, further comprising:
2 calculating the corpus weight as a function of a reference count of the at
3 least one concept over the plurality of documents.

1 15. (currently amended): A method according to Claim 9, further
2 comprising:
3 forming the score assigned to the at least one concept to a normalized
4 score vector for each such document;
5 determining [[a]] each such similarity between the normalized score
6 vector for each such document as an inner product of each normalized score
7 vector; and
8 applying the similarity to the best fit criterion.

1 16. (original): A method according to Claim 9, further comprising:
2 evaluating a set of candidate seed documents selected from the plurality of
3 documents;
4 identifying a set of seed documents by applying the score for the at least
5 one concept to a best fit criterion for each such candidate seed document; and
6 basing the best fit criterion on the score of each such seed document.

1 17. (currently amended): A computer-readable storage medium
2 holding code for grouping clusters of semantically scored documents
3 electronically stored in a data corpus, comprising:
4 code for determining a score, which is assigned to at least one concept that
5 has been extracted from a plurality of electronically-stored documents, wherein
6 the score is based on at least one of a frequency of occurrence of the at least one
7 concept within at least one such document, a concept weight, a structural weight,
8 and a corpus weight;
9 code for forming logically-grouped clusters of the documents by
10 evaluating the score for the at least one concept of each document for a best fit to
11 the clusters and assigning each document to the cluster with the best fit;

12 code for determining similarities between the documents grouped into
13 each cluster based on the center of the cluster and the scores assigned to each of
14 the at least one concepts in each such document;

15 code for dynamically determining a threshold for each cluster ~~based on as~~
16 a function of the similarities between the documents grouped into the cluster and
17 ~~a center of the cluster;~~ and

18 code for identifying and reassigning those documents having the
19 similarities falling outside the threshold.

1 18. (currently amended): A system for providing efficient document
2 scoring of concepts within and clustering of documents in an electronically-stored
3 document set, comprising:

4 a scoring module scoring a document in an electronically-stored document
5 set, comprising:

6 a frequency module determining a frequency of occurrence of at
7 least one concept within a document;

8 a concept weight module analyzing a concept weight reflecting a
9 specificity of meaning for the at least one concept within the document;

10 a structural weight module analyzing a structural weight reflecting
11 a degree of significance based on structural location within the document for the
12 at least one concept;

13 a corpus weight module analyzing a corpus weight inversely
14 weighing a reference count of occurrences for the at least one concept within the
15 document; and

16 a scoring evaluation module evaluating a score to be associated
17 with the at least one concept as a function of the frequency, concept weight,
18 structural weight, and corpus weight; and

19 a clustering module grouping the documents by score into a plurality of
20 clusters, comprising:

21 a cluster seed module identifying candidate seed documents, which
22 are each assigned as a seed document into a cluster with a center most similar to

23 the seed document, and assigning each non-seed document to the cluster with the
24 best fit; and
25 a threshold module relocating outlier documents, comprising
26 determining similarities between the documents grouped into each cluster based
27 on the center of the cluster and the scores assigned to each of the at least one
28 concepts in each such document, dynamically determining a threshold for each
29 cluster ~~based on as a function of the similarities between the documents grouped~~
30 ~~into the cluster and a center of the cluster~~, and identifying and reassigning the
31 documents with the similarities falling outside the threshold.

1 19. (previously presented): A system according to Claim 18, further
2 comprising:
3 the scoring module evaluating the score in accordance with the formula:

$$4 \quad S_i = \sum_{j=1}^n f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

5 where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises
6 the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$
7 comprises the corpus weight for occurrence j of concept i .

1 20. (previously presented): A system according to Claim 19, further
2 comprising:
3 the concept weight module evaluating the concept weight in accordance
4 with the formula:

$$5 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

6 where cw_{ij} comprises the concept weight and t_{ij} comprises a number of terms for
7 occurrence j of each such concept i .

1 21. (previously presented): A system according to Claim 19, further
2 comprising:

3 the structural weight module evaluating the structural weight in
4 accordance with the formula:

$$5 \quad sw_{ij} = \begin{cases} 1.0, & \text{if}(j \approx \textit{SUBJECT}) \\ 0.8, & \text{if}(j \approx \textit{HEADING}) \\ 0.7, & \text{if}(j \approx \textit{SUMMARY}) \\ 0.5 & \text{if}(j \approx \textit{BODY}) \\ 0.1 & \text{if}(j \approx \textit{SIGNATURE}) \end{cases}$$

6 where sw_{ij} comprises the structural weight for occurrence j of each such concept i .

1 22. (previously presented): A system according to Claim 19, further
2 comprising:
3 the corpus weight module evaluating the corpus weight in accordance with
4 the formula:

$$5 \quad rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

6 where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for
7 occurrence j of each such concept i , T comprises a total number of reference
8 counts of documents in the document set, and M comprises a maximum reference
9 count of documents in the document set.

1 23. (previously presented): A system according to Claim 19, further
2 comprising:
3 a compression module compressing the score in accordance with the
4 formula:

$$5 \quad S'_i = \log(S_i + 1)$$

6 where S'_i comprises the compressed score for each such concept i .

1 24. (original): A system according to Claim 18, further comprising:
2 a global stop concept vector cache maintaining concepts and terms; and

3 a filtering module filtering selection of the at least one concept based on
4 the concepts and terms maintained in the global stop concept vector cache.

1 25. (original): A system according to Claim 18, further comprising:
2 a parsing module identifying terms within at least one document in the
3 document set, and combining the identified terms into one or more of the
4 concepts.

1 26. (original): A system according to Claim 25, further comprising:
2 the parsing module structuring each such identified term in the one or
3 more concepts into canonical concepts comprising at least one of word root,
4 character case, and word ordering.

1 27. (original): A system according to Claim 25, wherein at least one of
2 nouns, proper nouns and adjectives are included as terms.

1 28. (original): A system according to Claim 18, further comprising:
2 a plurality of candidate seed documents;
3 a similarity module determining a similarity between each pair of a
4 candidate seed document and a cluster center;
5 a clustering module designating each such candidate seed document
6 separated from substantially all cluster centers with such similarity being
7 sufficiently distinct as a seed document, and grouping each such candidate seed
8 document not being sufficiently distinct into a cluster with a nearest cluster
9 center.

1 29. (original): A system according to Claim 28, further comprising:
2 a plurality of non-seed documents;
3 the similarity module determining the similarity between each non-seed
4 document and each cluster center; and
5 the clustering module grouping each such non-seed document into a
6 cluster having a best fit, subject to a minimum fit criterion.

1 30. (original): A system according to Claim 29, further comprising:
2 a normalized score vector for each document comprising the score
3 associated with the at least one concept for each such concept occurring within
4 the document; and
5 the similarity module determining the similarity as a function of the
6 normalized score vector associated with the at least one concept for each such
7 document.

1 31. (previously presented): A system according to Claim 30, further
2 comprising:
3 the similarity module calculating the similarity in accordance with the
4 formula:

5
$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

6 where $\cos \sigma_{AB}$ comprises a similarity between a document A and a document B ,
7 \vec{S}_A comprises a score vector for document A , and \vec{S}_B comprises a score vector for
8 document B .

1 Claims 32-34 (canceled).

1 35. (currently amended): A method for providing efficient document
2 scoring of concepts within and clustering of documents in an electronically-stored
3 document set, comprising:
4 scoring a document in an electronically-stored document set, comprising:
5 determining a frequency of occurrence of at least one concept
6 within a document;
7 analyzing a concept weight reflecting a specificity of meaning for
8 the at least one concept within the document;
9 analyzing a structural weight reflecting a degree of significance
10 based on structural location within the document for the at least one concept;

11 analyzing a corpus weight inversely weighing a reference count of
12 occurrences for the at least one concept within the document; and
13 evaluating a score to be associated with the at least one concept as
14 a function of the frequency, concept weight, structural weight, and corpus weight;
15 and
16 grouping the documents by score into a plurality of clusters, comprising:
17 identifying candidate seed documents, which are each assigned as
18 a seed document into a cluster with a center most similar to the seed document;
19 assigning each non-seed document to the cluster with the best fit;
20 relocating outlier documents, comprising:
21 determining similarities between the documents grouped into each
22 cluster based on the center of the cluster and the scores assigned to each of the at
23 least one concepts in each such document;
24 dynamically determining a threshold for each cluster based on as a
25 function of the similarities between the documents grouped into the cluster and a
26 center of the cluster; and
27 identifying and reassigning the documents with the similarities
28 falling outside the threshold.

1 36. (previously presented): A method according to Claim 35, further
2 comprising:

3 evaluating the score in accordance with the formula:

4
$$S_i = \sum_{l \rightarrow n}^j f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

5 where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises
6 the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$
7 comprises the corpus weight for occurrence j of concept i .

1 37. (previously presented): A method according to Claim 36, further
2 comprising:

3 evaluating the concept weight in accordance with the formula:

$$4 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

5 where cw_{ij} comprises the concept weight and t_{ij} comprises a number of terms for
6 occurrence j of each such concept i .

1 38. (previously presented): A method according to Claim 36, further
2 comprising:
3 evaluating the structural weight in accordance with the formula:

$$4 \quad sw_{ij} = \begin{cases} 1.0, & \text{if}(j \approx \text{SUBJECT}) \\ 0.8, & \text{if}(j \approx \text{HEADING}) \\ 0.7, & \text{if}(j \approx \text{SUMMARY}) \\ 0.5 & \text{if}(j \approx \text{BODY}) \\ 0.1 & \text{if}(j \approx \text{SIGNATURE}) \end{cases}$$

5 where sw_{ij} comprises the structural weight for occurrence j of each such concept i .

1 39. (previously presented): A method according to Claim 36, further
2 comprising:
3 evaluating the corpus weight in accordance with the formula:

$$4 \quad rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

5 where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for
6 occurrence j of each such concept i , T comprises a total number of reference
7 counts of documents in the document set, and M comprises a maximum reference
8 count of documents in the document set.

1 40. (previously presented): A method according to Claim 36, further
2 comprising:
3 compressing the score in accordance with the formula:

4 $S'_i = \log(S_i + 1)$

5 where S'_i comprises the compressed score for each such concept i .

1 41. (original): A method according to Claim 35, further comprising:
2 maintaining concepts and terms in a global stop concept vector cache; and
3 filtering selection of the at least one concept based on the concepts and
4 terms maintained in the global stop concept vector cache.

1 42. (original): A method according to Claim 35, further comprising:
2 identifying terms within at least one document in the document set; and
3 combining the identified terms into one or more of the concepts.

1 43. (original): A method according to Claim 42, further comprising:
2 structuring each such identified term in the one or more concepts into
3 canonical concepts comprising at least one of word root, character case, and word
4 ordering.

1 44. (original): A method according to Claim 42, further comprising:
2 including as terms at least one of nouns, proper nouns and adjectives.

1 Claim 45 (canceled).

1 46. (previously presented): A method according to Claim 35, further
2 comprising:
3 identifying a plurality of non-seed documents;
4 determining the similarity between each non-seed document and each
5 cluster center; and
6 grouping each such non-seed document into a cluster with a best fit,
7 subject to a minimum fit criterion.

1 47. (original): A method according to Claim 46, further comprising:

2 forming a normalized score vector for each document comprising the
3 score associated with the at least one concept for each such concept occurring
4 within the document; and
5 determining the similarity as a function of the normalized score vector
6 associated with the at least one concept for each such document.

1 48. (previously presented): A method according to Claim 47, further
2 comprising:
3 calculating the similarity in accordance with the formula:

4
$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

5 where $\cos \sigma_{AB}$ comprises a similarity between a document A and a document B ,
6 \vec{S}_A comprises a score vector for document A , and \vec{S}_B comprises a score vector for
7 document B .

1 Claims 49-51 (canceled).

1 52. (currently amended): A computer-readable storage medium
2 holding code for providing efficient document scoring of concepts within and
3 clustering of documents in an electronically-stored document set, comprising:
4 code for scoring a document in an electronically-stored document set,
5 comprising:
6 code for determining a frequency of occurrence of at least one
7 concept within a document;
8 code for analyzing a concept weight reflecting a specificity of
9 meaning for the at least one concept within the document;
10 code for analyzing a structural weight reflecting a degree of
11 significance based on structural location within the document for the at least one
12 concept;
13 code for analyzing a corpus weight inversely weighing a reference
14 count of occurrences for the at least one concept within the document; and

15 code for evaluating a score to be associated with the at least one
16 concept as a function of the frequency, concept weight, structural weight, and
17 corpus weight; and
18 code for grouping the documents by score into a plurality of clusters,
19 comprising:
20 code for identifying candidate seed documents, which are each
21 assigned as a seed document into a cluster with a center most similar to the seed
22 document;
23 code for assigning each non-seed document to the cluster with the
24 best fit;
25 code for relocating outlier documents, comprising:
26 code for determining similarities between the documents grouped
27 into each cluster based on the center of the cluster and the scores assigned to each
28 of the at least one concepts in each such document;
29 code for dynamically determining a threshold for each cluster
30 ~~based on as a function of the similarities between the documents grouped into the~~
31 ~~cluster and a center of the cluster;~~ and
32 code for identifying and reassigning the documents with the
33 similarities falling outside the threshold.

1 53. (currently amended): An apparatus for providing efficient
2 document scoring of concepts within and clustering of documents in an
3 electronically-stored document set, comprising:
4 means for scoring a document in an electronically-stored document set,
5 comprising:
6 means for determining a frequency of occurrence of at least one
7 concept within a document;
8 means for analyzing a concept weight reflecting a specificity of
9 meaning for the at least one concept within the document;

10 means for analyzing a structural weight reflecting a degree of
11 significance based on structural location within the document for the at least one
12 concept;

13 means for analyzing a corpus weight inversely weighing a
14 reference count of occurrences for the at least one concept within the document;
15 and

16 means for evaluating a score to be associated with the at least one
17 concept as a function of the frequency, concept weight, structural weight, and
18 corpus weight; and

19 means for grouping the documents by score into a plurality of clusters,
20 comprising:

21 means for identifying candidate seed documents, which are each
22 assigned as a seed document into a cluster with a center most similar to the seed
23 document;

24 means for assigning each non-seed document to the cluster with
25 the best fit;

26 means for relocating outlier documents, comprising:

27 means for determining similarities between the documents grouped
28 into each cluster based on the center of the cluster and the scores assigned to each
29 of the at least one concepts in each such document;

30 means for dynamically determining a threshold for each cluster
31 ~~based on as a function of the similarities between the documents grouped into the~~
32 ~~cluster and a center of the cluster;~~ and

33 means for identifying and reassigning the documents with the
34 similarities falling outside the threshold.